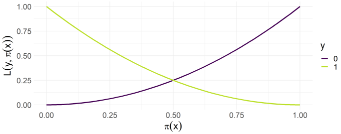
Introduction to Machine Learning

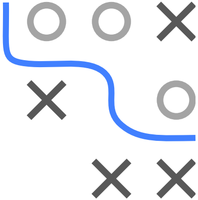
Brier Score



Learning goals

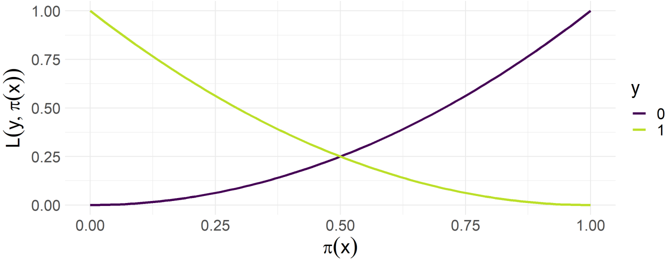
- Know the Brier score
- Derive the risk minimizer
- Derive the optimal constant model
- Understand the connection between Brier score and Gini splitting

BRIER SCORE

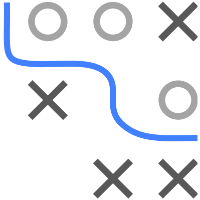


The binary Brier score is defined on probabilities $\pi(\mathbf{x}) \in [0, 1]$ and 0-1-encoded labels $y \in \{0, 1\}$ and measures their squared distance ($L2$ loss on probabilities).

$$L(y, \pi(\mathbf{x})) = (\pi(\mathbf{x}) - y)^2$$



BRIER SCORE: RISK MINIMIZER



The risk minimizer for the (binary) Brier score is

$$\pi^*(\mathbf{x}) \quad = \quad \eta(\mathbf{x}) = \mathbb{P}(y \mid \mathbf{x} = \mathbf{x}),$$

which means that the Brier score will reach its minimum if the prediction equals the “true” probability of the outcome.

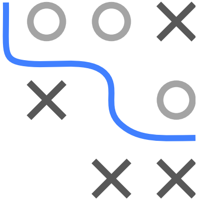
The risk minimizer for the multiclass Brier score is

$$\pi^*(\mathbf{x}) = \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}).$$

Proof: We only show the proof for the binary case. We need to minimize

$$\mathbb{E}_x \left[L(1, \pi(\mathbf{x})) \cdot \eta(\mathbf{x}) + L(0, \pi(\mathbf{x})) \cdot (1 - \eta(\mathbf{x}))\right],$$

BRIER SCORE: RISK MINIMIZER / 2

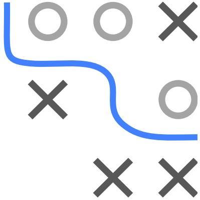


which we do point-wise for every $\mathbf{x}$. We plug in the Brier score

$$\begin{aligned} & \arg\min_c L(1, c)\eta(\mathbf{x}) + L(0, c)(1 - \eta(\mathbf{x})) \\ = \quad & \arg\min_c \quad (c-1)^2\eta(\mathbf{x}) + c^2(1 - \eta(\mathbf{x})) \\ = \quad & \arg\min_c \quad (c - \eta(\mathbf{x}))^2. \end{aligned}$$

The expression is minimal if $c = \eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$.

BRIER SCORE: OPTIMAL CONSTANT MODEL



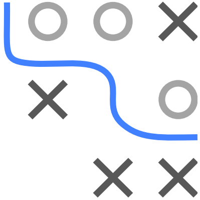
The optimal constant probability model $\pi(\mathbf{x}) = \theta$ w.r.t. the Brier score for labels from $\mathcal{Y} = \{0, 1\}$ is:

$$
\begin{aligned}
\min_{\theta} \mathcal{R}_{\text{emp}}(\theta) &= \min_{\theta} \sum_{i=1}^{n} \left(y^{(i)} - \theta\right)^2 \\
\Leftrightarrow \frac{\partial \mathcal{R}_{\text{emp}}(\theta)}{\partial \theta} &= -2 \cdot \sum_{i=1}^{n} (y^{(i)} - \theta) = 0 \\
\hat{\theta} &= \frac{1}{n} \sum_{i=1}^{n} y^{(i)}.
\end{aligned}
$$

This is the fraction of class-1 observations in the observed data.
(This also directly follows from our $L2$ proof for regression).

Similarly, for the multiclass brier score the optimal constant is
$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} [y = k]$.

BRIER SCORE MINIMIZATION = GINI SPLITTING



When fitting a tree we minimize the risk within each node $\mathcal{N}$ by risk minimization and predict the optimal constant. Another approach that is common in literature is to minimize the average node impurity $\text{Imp}(\mathcal{N})$.

Claim: Gini splitting $\text{Imp}(\mathcal{N}) = \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \left(1 - \pi_k^{(\mathcal{N})}\right)$ is equivalent to the Brier score minimization.

Note that $\pi_k^{(\mathcal{N})} := \frac{1}{n_{\mathcal{N}}} \sum\limits_{(\mathbf{x},y)\in\mathcal{N}} [y = k]$

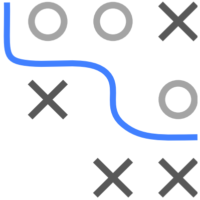
Proof: We show that the risk related to a subset of observations $\mathcal{N} \subseteq \mathcal{D}$ fulfills

$$\mathcal{R}(\mathcal{N}) = n_{\mathcal{N}} \text{Imp}(\mathcal{N}),$$

where Imp is the Gini impurity and $\mathcal{R}(\mathcal{N})$ is calculated w.r.t. the (multiclass) Brier score

$$L(y, \pi(\mathbf{x})) = \sum_{k=1}^{g} \left([y = k] - \pi_k(\mathbf{x})\right)^2 .$$

BRIER SCORE MINIMIZATION = GINI SPLITTING / 2



$$\mathcal{R}(\mathcal{N}) \quad = \quad \sum_{(\mathbf{x},y)\in\mathcal{N}} \sum_{k=1}^{g} ([y=k] - \pi_k(\mathbf{x}))^2 = \sum_{k=1}^{g} \sum_{(\mathbf{x},y)\in\mathcal{N}} \left([y=k] - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2,$$

by plugging in the optimal constant prediction w.r.t. the Brier score ($n_{\mathcal{N},k}$ is defined as the number of class k observations in node $\mathcal{N}$):

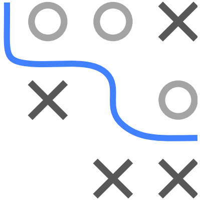
$$\hat{\pi}_k(\mathbf{x}) = \pi_k^{(\mathcal{N})} = \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x},y)\in\mathcal{N}} [y=k] = \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}.$$

We split the inner sum and further simplify the expression

$$\begin{aligned}
&= \sum_{k=1}^{g} \left(\sum_{(\mathbf{x},y)\in\mathcal{N}:\ y=k} \left(1 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 + \sum_{(\mathbf{x},y)\in\mathcal{N}:\ y\neq k} \left(0 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 \right) \\
&= \sum_{k=1}^{g} n_{\mathcal{N},k} \left(1 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k}) \left(\frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2,
\end{aligned}$$

since for $n_{\mathcal{N},k}$ observations the condition $y=k$ is met, and for the remaining $(n_{\mathcal{N}} - n_{\mathcal{N},k})$ observations it is not.

BRIER SCORE MINIMIZATION = GINI SPLITTING / 3



We further simplify the expression to

$$
\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{k=1}^{g} n_{\mathcal{N},k} \left(\frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}} \right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k}) \left(\frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}} \right)^2 \\
&= \sum_{k=1}^{g} \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}} \frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}} (n_{\mathcal{N}} - n_{\mathcal{N},k} + n_{\mathcal{N},k}) \\
&= n_{\mathcal{N}} \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \cdot \left(1 - \pi_k^{(\mathcal{N})} \right) = n_{\mathcal{N}} \mathrm{Imp}(\mathcal{N}).
\end{aligned}
$$